Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

I have identified, season and weathersit as categorical variables

If I use those variables as is, as those are not numerical, I can use those directly in regression. Instead of that I must convert those to dummy variables and creating those variables help to convert it into 1 or 0 and use those in the regression.

Winter has better coefficient than fall for the dependent variable cnt

Where as weathershot has negative coefficient for the cnt

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop_first=True during dummy variable creation is important because it helps to avoid the dummy variable trap, which occurs when there is multicollinearity among the dummy variables. Multicollinearity can lead to issues in regression models, making it difficult to interpret the coefficients and potentially leading to unreliable estimates. When you have a categorical variable with n levels, creating n dummy variables would result in perfect multicollinearity because the sum of all dummy variables would always equal 1. By setting drop_first=True, you drop the first dummy variable, which reduces the number of dummy variables to n-1. This way, you avoid multicollinearity while still capturing all the information about the categorical variable.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Temp variable had the highest correlation with the cnt variable

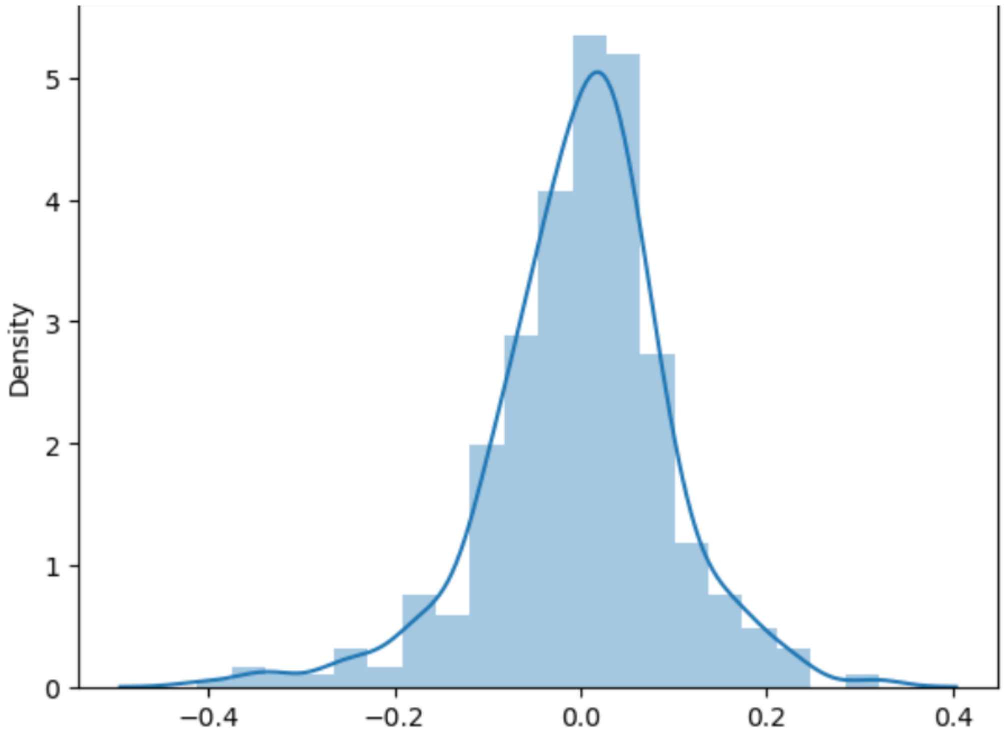
---

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Normality of Residuals: checked if the residuals are normally distributed. Plotted a histogram of the residuals. As the residuals follow a normal distribution, the histogram is bell-shaped



Multicollinearity: Ensured that the independent variables are not highly correlated with each other. This can be checked using the Variance Inflation Factor (VIF). All VIF value are less than 5 .

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

Temp, yr and winter

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear Regression Algorithm Explained

Linear regression is a fundamental algorithm in machine learning and statistics. It's used to model the relationship between a dependent variable (often called the target or output) and one or more independent variables (often called features or inputs).

Key Concepts

1. Dependent and Independent Variables :
   - Dependent Variable (Y) : The outcome we are trying to predict.
   - Independent Variables (X) : The features or inputs that we use to make predictions.

2. Linear Relationship :
   - Linear regression assumes a linear relationship between the dependent and independent variables. This means that the change in the dependent variable is proportional to the change in the independent variable(s).

The Linear Regression Equation

The equation for a simple linear regression (with one independent variable) is:

$$ Y = \beta_0 + \beta_1X + \epsilon $$

- Y : Dependent variable (what we are predicting)
- X : Independent variable (feature)
- \beta_0 : Intercept (the value of Y when X is 0)
- \beta_1 : Slope (the change in Y for a one-unit change in X)
- \epsilon : Error term (the difference between the actual and predicted values)

Steps to Perform Linear Regression

1. Data Collection :
   - Gather data with known values of the dependent and independent variables.

2. Data Preprocessing :
   - Clean the data, handle missing values, and possibly normalize or standardize the features.

3. Model Training :
   - Use the training data to estimate the coefficients (\(\beta_0\) and \(\beta_1\)) that minimize the error term (\(\epsilon\)).

4. Model Evaluation :
   - Evaluate the model's performance using metrics like Mean Squared Error (MSE) or R-squared.

5. Prediction :
   - Use the trained model to make predictions on new data.

Example

Let's say we want to predict a student's exam score (Y) based on the number of hours they studied (X). We collect data from several students and fit a linear regression model. The resulting equation might look like this:

$$ \text{Exam Score} = 50 + 5 \times (\text{Hours Studied}) $$

This means that for every additional hour studied, the exam score increases by 5 points, starting from a base score of 50.

Conclusion

Linear regression is a powerful and simple algorithm for predicting continuous outcomes. By understanding the relationship between variables, we can make informed predictions and gain insights into the data.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>
Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and to show how statistical properties can be misleading if not visualized.

Key Characteristics

Each of the four datasets in Anscombe's quartet has the following nearly identical statistical properties:

- Mean of x : 9
- Mean of y : 7.5
- Variance of x : 11
- Variance of y : 4.125
- Correlation between x and y : 0.816
- Linear regression line : \(y = 3 + 0.5x \)
- Coefficient of determination (R^2) : 0.67

Importance of Anscombe's Quartet

Anscombe's quartet highlights several important lessons in data analysis:

1. Visualization : Always visualize your data. Graphs can reveal patterns, relationships, and anomalies that summary statistics might miss.
2. Outliers : Outliers can significantly affect statistical measures and regression models. Identifying and understanding outliers is crucial.
3. Context : Statistical properties alone do not provide a complete picture. The context and distribution of the data are essential for accurate interpretation.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>
Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear relationship between two variables. It quantifies the strength and direction of this relationship,

providing a value between -1 and 1.

Key Points

1. Value Range :
   - +1 : Perfect positive linear relationship.
   - 0 : No linear relationship.
   - -1 : Perfect negative linear relationship.

2. Interpretation :
   - Positive Values : As one variable increases, the other variable also increases.
   - Negative Values : As one variable increases, the other variable decreases.
   - Closer to ±1 : Stronger linear relationship.
   - Closer to 0 : Weaker linear relationship.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>
What is Scaling?

Scaling is the process of transforming the features of your data so that they are on a similar scale. This is particularly important in machine learning algorithms that are sensitive to the magnitude of the features, such as gradient descent-based algorithms and distance-based algorithms like k-nearest neighbors (KNN).

Why is Scaling Performed?

Scaling is performed to:
1. Improve Model Performance : Algorithms like gradient descent converge faster when features are on a similar scale.
2. Enhance Accuracy : Distance-based algorithms (e.g., KNN, SVM) perform better when features are scaled, as they rely on the distance between data points.
3. Prevent Dominance : Features with larger magnitudes can dominate the learning process, leading to biased models.

Normalized Scaling vs. Standardized Scaling

Normalized Scaling

- Definition : Normalization (or Min-Max Scaling) transforms the data to fit within a specific range, usually [0, 1].
- Formula :

$$ X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}} $$

- Use Case : Useful when you want to bound the feature values within a specific range. It's commonly used in algorithms that do not assume any distribution of the data.

Standardized Scaling

- Definition : Standardization (or Z-score Scaling) transforms the data to have a mean of 0 and a standard deviation of 1.
- Formula :

$$ X' = \frac{X - \mu}{\sigma} $$

Where \(\mu \) is the mean and \(\sigma \) is the standard deviation of the feature.

- Use Case : Useful when the data follows a Gaussian distribution. It's commonly used in algorithms that assume the data is normally distributed.

Summary

- Normalization scales the data to a fixed range, typically [0, 1].
- Standardization scales the data to have a mean of 0 and a standard deviation of 1.
- Both methods help improve the performance and accuracy of machine learning models by ensuring that features contribute equally to the learning process.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>
An infinite value of the Variance Inflation Factor (VIF) indicates perfect multicollinearity. This happens when one independent variable in a regression model can be perfectly predicted by a linear combination of the other independent variables. In other words, the variable is completely redundant because it does not provide any new information beyond what is already explained by the other variables.

Mathematically, this occurs when the R^2 value in the VIF formula approaches 1:

$$ \text{VIF} = \frac{1}{1 - R^2} $$

When \(R^2 \) is very close to 1, the denominator approaches zero, causing the VIF to approach infinity

---

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>
A Q-Q plot, or Quantile-Quantile plot, is a graphical tool used to compare the distribution of a dataset to a theoretical distribution, typically the normal distribution. It helps to assess whether the data follows a particular distribution.

What is a Q-Q Plot?

- Axes : The x-axis represents the theoretical quantiles, and the y-axis represents the sample quantiles.
- Line of Best Fit : If the data follows the theoretical distribution, the points will approximately lie on a straight line (the line of best fit).

Use and Importance in Linear Regression

1. Assessing Normality :
   - In linear regression, one of the assumptions is that the residuals (errors) are normally distributed. A Q-Q plot helps to check this assumption.
   - If the residuals are normally distributed, the points on the Q-Q plot will lie on a straight line.

2. Identifying Deviations :
   - Deviations from the straight line indicate departures from normality. For example:
     - Heavy Tails : Points that deviate upwards or downwards at the ends suggest heavy tails.
     - Skewness : Points that curve away from the line suggest skewness in the data.

3. Model Validation :
   - By using a Q-Q plot to check the normality of residuals, you can validate the assumptions of your linear regression model. This helps ensure the reliability and accuracy of your model's predictions.